Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity

Lars Backstrom
lars@facebook.com

Eric Sun
esun@facebook.com

Cameron Marlow
cameron@facebook.com

1601 S. California Ave.
Palo Alto, CA 94304

ABSTRACT

Geography and social relationships are inextricably intertwined; the people we interact with on a daily basis almost always live near us. As people spend more time online, data regarding these two dimensions – geography and social relationships – are becoming increasingly precise, allowing us to build reliable models to describe their interaction. These models have important implications in the design of location-based services, security intrusion detection, and social media supporting local communities.

Using user-supplied address data and the network of associations between members of the Facebook social network, we can directly observe and measure the relationship between geography and friendship. Using these measurements, we introduce an algorithm that predicts the location of an individual from a sparse set of located users with performance that exceeds IP-based geolocation. This algorithm is efficient and scalable, and could be run on a network containing hundreds of millions of users.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Measurement, Theory

Keywords

social networks, geolocation, propagation

1. INTRODUCTION

While we would like to believe that our social options are endless, human relationships are constrained in many ways. They take time, energy, and often money to maintain. Even after accounting for these human constraints, social norms dictate whom we approach and how we become acquainted. All of these constraints create a predictable structure where geography, transportation, employment, and existing relationships predict the set of people with whom we will associate and communicate.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

We have long observed that the likelihood of friendship with a person is decreasing with distance. This should not be surprising given that we are less likely to meet people who live further away. A less obvious relationship [18] is that the total number of friends also tends to decrease as distance increases. This means that the probability of knowing someone d miles away is decreasing faster than the total number of people d miles away is increasing.

The Internet and other communication technologies play a potentially disruptive role on the constraints imposed on social networks. These technologies reduce the overhead and cost for being introduced to new people regardless of geography, and help us stay in touch with those we know. Some have even gone so far as to call this "the end of geography," where the process of relationship formation becomes disentangled from distance altogether [9]. As people conduct more and more of their lives online, data about location and social relationships become increasingly precise. While geography is certainly playing a smaller role in our lives than it once did, we see in this work that geography is far from over.

Geography has a number of compelling applications within Internet technology, and accurately predicting a user's location can significantly improve a user's experience. First, as malicious entities create increasingly compelling "phishing" sites that deceive users into providing their account credentials [6], it becomes difficult to identify when an account has been compromised. Having a good baseline understanding of a user's geography along with IP geolocation allows for the detection of masquerading accounts. Second, knowing a user's general location can allow for personalization based on location. Instead of requiring a user to specify information about themselves, a news site can immediately provide local stories or an international service can set the default language of the interface automatically.

The current industry standard for geolocation depends on mapping a user's IP address to a known or predicted location, and these services typically provide accuracy at the city level. However, results are inconsistent – for example, customers of large or mobile Internet service providers are generally assigned IP addresses from large pools, rendering accurate geolocation difficult. These inconsistencies spill over into the user experience; nothing is more jarring than having your default language switched to French because a service has incorrectly determined that your IP address is in France. While we do not have the capability to evaluate the performance of the many IP location services available, a leading service reports their accuracy as locating 85% of US

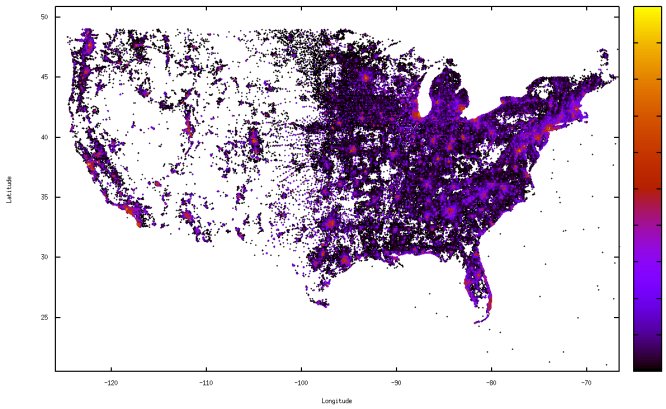

Figure 1: US population density of geolocated Facebook users.

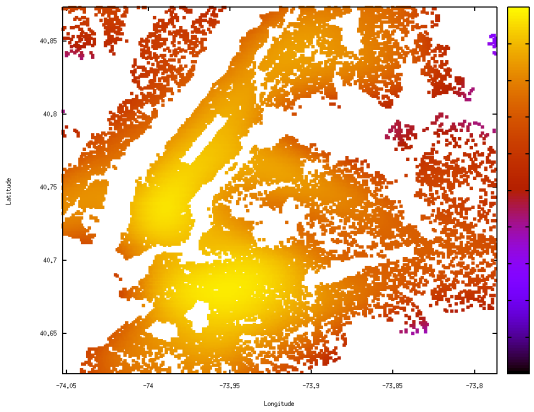

Figure 2: NY population density of geolocated Facebook users.

IPs within 25 miles, with performance of only 59% in the UK [16]. Furthermore, maintainance of this performance requires constant updating as IPs are reassigned and performance drops about 1.5% per month without this effort.

In this paper we present the findings of a large-scale study on social and spatial proximity using relationships expressed on the Facebook social network between users in the United States. First, we examine the relationship between proximity and friendship, observing that, as expected, the likelihood of friendship drops monotonically as a function of distance. This effect can also be seen as a function of rank, where friendships are assumed to be independent of their explicit distance. Second, we use this distribution to derive the maximum-likelihood location of individuals with unknown location and show that this model outperforms data provided by geolocation services based on a person's IP address. Finally, we introduce an iterative algorithm to refine our predictions based on the propagation of predictions across the network holding out a large percentage of known locations for evaluation purposes. In all of our analyses, data were anonymized and analyzed in aggregate so as to ensure the privacy of our users.

While other geolocation strategies depend on opaque mappings between proprietary databases and geography, the techniques provided in this paper use entirely transparent methods which are easily understood by users. By deriving a user's location through friends' geography, we can take advantage of all of the affordances of location-enabled services without the obscurity and data errors in existing systems.

Our contributions are thus twofold. First, the number of individuals and the precision to which we can locate them allows us to study the interplay between geographic distance and social relationship with greater accuracy and in greater depth than has previously been possible. Second, using some of our observations concerning this interplay, we are able to develop algorithms which locate users with greater accuracy than existing IP-based methods. Not only does this improve our accuracy when it comes to various locality related tasks, but it also mitigates the need for constant maintainance of geo-IP databases.

2. BACKGROUND

In this section we review the empirical and theoretical work that informs the central questions of this paper: how does geography bound social structure, and in what ways can this relationship inform location prediction?

Sociologists and social psychologists have long studied the relationship between propinquity and friendship. The geography and social environment that one experiences largely dictates the people and information that one has access to. Over the years, many researchers have noted an inverse relationship between distance and the likelihood of friendship. This has been expressed simply as a decrease in the probability of coming into contact with one another, and has been observed within colleges [20], new housing developments [7], and projects for the elderly [19]. In addition to affecting the likelihood of friendship, density and spatial arrangement of people is expected to have an impact on the size and frequency of interaction among social ties [17]. These observations seem to hold across time, technological innovation, and culture, although recent changes in technology are changing the way that relationships persist over time [18].

The Internet brings both a potential to disrupt the relationship between distance and friendship as well as to introduce unprecedented data validating these theories at the level of an entire human population. From the analysis of early social networking technology [1] to the whole-network analyses performed on entire communities of users, such as LiveJournal, LinkedIn, and Flickr [2, 13, 14], social media and social networking communities are nearing the scale of entire countries. The level of transactional detail afforded by these services allows for analysis and modeling that bridges micro-level processes and population-level effects.

Recently, the question of propinquity and social structure has been at the center of research around routing in small-world networks [11, 12]. Using the networks and cities of US LiveJournal members, Liben-Nowell et al. observe a number of properties of geographic and social proximity [15]. Most notably, they find that the likelihood of friendship is inversely proportional to distance, but at extremely long distances, there is a baseline probability of geographic-independent relationships that takes over. To account for the confounding effects of population density, they introduce the notion of rank-based distance, measuring the probability

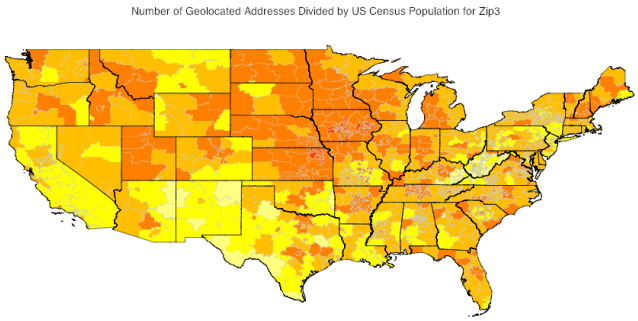

Figure 3: ***Facebook penetration using user-provided addresses.*** **As a proportion of population, users in the midwest share more addresses on Facebook. However, this corresponds closely to overall Facebook penetration, shown in the next figure.**

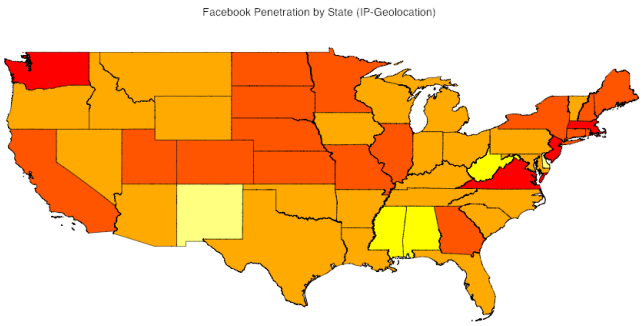

Figure 4: ***Facebook penetration using IP geolocation.*** **Facebook penetration by state, normalized by each state's population.**

that v will be u's friend given the number of people w such that $dist_{u,w} < dist_{u,v}$.

In another recent analysis of a large sample of MySpace users in the United States, Gilbert et al. studied differences in behavior between urban and rural users [8]. Dividing relationships into strong and weak ties based on communication frequency, they found that urban users' ties and strong ties tended to be more geographically distributed than rural and weak ties. While the distances and lack of scale-invariance disagree with Liben-Nowell et al., these results show continued evidence for an inverse relationship between distance and acquaintanceship within the US population.

On the non-social front, there has been an increasing interest in geographic properties. In [3], Yahoo! search queries were used in the development of an algorithm that accurately located the geographic center and rate of diffusion for various query terms. Despite the sparsity of some queries, such as "Grand Canyon National Park," this algorithm was able to correctly position the center of the query to only about 50 miles from the actual park. In a study of the Flickr photo-sharing website [5], Crandall et al. were able to automatically locate landmarks based on geotagged photos. Furthermore, once the location was identified they presented an algorithm which extracted representative images of the landmark at that location using photographic content. These studies illustrate the practicality of meaningful geographic work on these sorts of large, noisy, user-generated datasets.

Although propinquity and friendship has been a topic of study across many decades and disciplines, the observations of the earliest studies have not changed: the further you get from a person, the lower the likelihood you'll find her friends there. Most of the literature has focused on using geography to explain and model relationships [4], and in this paper we would like to propose the reverse: given a set of relationships and some knowledge of geography, how well can we predict the location of others in the network? The remainder of this paper is divided into three sections: first, we discuss descriptive properties of the Facebook network and geographic data, paying specific attention to density and friendship as a function of distance; second, we describe the use of these observations in a predictive model, along with a number of optimizations; finally, we conclude with applications and future work.

3. DATA

Of the roughly 100 million Facebook users in the United States, a small but significant fraction (about 6%) have elected to enter their home address. Of those who have entered their addresses, roughly 60% of the addresses can be easily parsed and converted to latitude and longitude using the publicly available TIGER/Line data set from the United States Census Bureau [22]. This gives us a set of approximately 3.5 million users with precisely known home addresses.

Naturally, some of these addresses are incorrect or out of date, but there is little incentive to enter false information, as leaving the field blank is an easier option. Furthermore, addresses that are ambiguous or do not include precise street numbers are ignored.

Of the 3.5 million users with addresses, 2.9 million also have at least one friend with a valid address, and on average they have 10 friends with addresses, giving us 30.6 million edges between individuals with known locations. Having so many edges between individuals whose addresses are known so precisely allows us to study the relationship between distance and friendship on a scale not previously possible, and with greater precision than in previous studies (which tended to garner location from IP addresses, which are inherently imprecise).

In order to uncover potential sources of bias in our methods and learn more about the users that choose to supply addresses on Facebook, we compare the demographic attributes of users who disclose their location to those who do not. Table 1 shows demographic statistics for the geolocated users compared to the overall Facebook population in the United States.

Users of different ages are roughly equally likely to share their address information. However, males are significantly more likely to share their address information than females. This agrees with many studies that show that males tend to share more personal data online [10]. Furthermore, users that share their addresses tend to have many more friends. This could be because these users also tend to be longer-tenured users of Facebook.

Table 1: Demographic Statistics of Geolocated Users

	Located	All US Users
% Male	57.51%	44.81%
% Female	42.49%	55.19%
Age, Median	30	30
Age, Mean	33.89	33.44
Account Age (days), Median	413	325
Account Age (days), Mean	558.9	453
Friend Count, Median	105	47
Friend Count, Mean	189.4	129.5

Since the the number of geolocated addresses in our data is a relatively small fraction of the overall US population, bias may also result if people in some parts of the country are more likely to share address information than others. To investigate this potential concern, Figure 3 shows a heatmap of the number of geolocated addresses divided by US Census population (from the 2000 US census) for each 3-digit ZIP code tabulation area (ZIP3) [21]. This does not cause great concern because the heatmap corresponds closely to Figure 4, which shows Facebook penetration by state using IP-based geolocation. Differences in certain states may be due to large pools of IP addresses owned by large Internet service providers.

3.1 Population Density

In order to understand the dynamics between population and geography, we first examine the distribution of density in our sample. We divide the United States into a cells of 1/100 of a degree square, or roughly 0.4 square miles in the continental US. Figure 5 shows the number of grid units in our data as a function of the density (number of people). Plotting on a log-log scale, we see that the curve has two regions. In the low density area, the distribution is decreasing roughly according to a power-law with exponent -1.37. At some point there is a transition into higher density region where the exponent decreases to -3.07. This transition occurs at about 50 people per square mile, or 560,000 square feet per person. Since this includes only Facebook members who have provided an address, we would expect the actual density at this transition point to be only about 5600 square feet per person – about the density of a densely populated suburban area. In fact, our data illustrates that 96% of people live in areas less dense than this, suggesting that the -1.37 exponent is the one which we should focus on, and that the distribution takes an abrupt downward turn as we transition into the density of large apartment complexes.

Figure 1 shows the distribution of the geolocated individuals across the United States. To smooth these figures, a Gaussian kernel has been applied to each individual, with width 1 mile. Some artifacts of the geolocation appear in the ocean, but are overrepresented by this visualization and account for a negligible fraction of all users. Note that the vast majority of the country is quite sparsely populated, and in fact about half of the US population lives in regions with less than 250 people per square mile (this is the scaled-up value which accounts for the fact that only 1% of the US population has provided us with geolocatable addresses). It is important to note, however, that this is somewhat biased

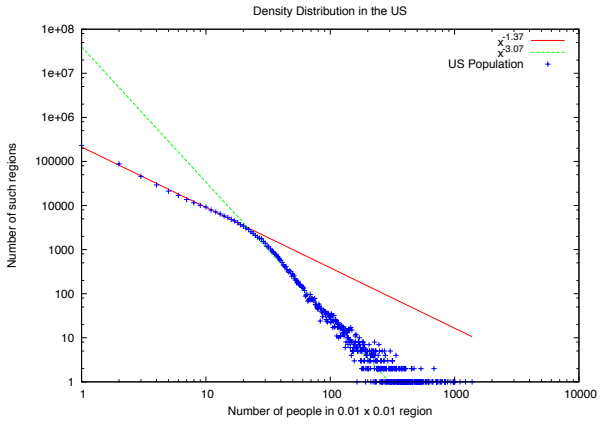

Figure 5: The density distribution of the US. The country is divided into 0.01 x 0.01 degree regions (about 0.4 square miles). We then count the number of Facebook members in each region and plot the distribution of counts. There seem to be two distinct regions of the distribution, a low-density region where the curve fits a straight line (in log-log space) with slope -1.37 and a high-density region, where the fall off is much sharper with slope -3.07.

by the differences in Facebook demographics as compared to the demographics of the US.

It has been observed in other contexts that the interplay between distance and friendship is in some way connected to population density. If you live in Manhattan and have thousands of people living within a single block, you are not particularly likely to know any one of them. For example, if you knew five out of ten thousand people within 1 mile, then your probability of knowing any one individual would only be 0.0005. Contrast this with a small town setting where everyone has a large yard and there are only a thousand people within a mile. In this case you might still only know five other people within a mile, but your probability for each person would be 0.005, an order of magnitude higher.

The first part of this relationship is shown in Figure 6. Here we divide the population of the United States into three groups of roughly equal size (about 900,000 people per group) according to the population density where they live. This figure shows the average number of people living x miles away, as a function of x. Note that this is not the number living within x miles, but is the number living within the annulus of width 0.1 miles.

By definition, there are more people living nearby in the high density case. If the population were uniformly distributed, we would expect the curves to increase linearly, since the area of an annulus with inner radius r and width w is $\pi((r+w)^2-r^2) = \pi(2rw+w^2)$, roughly linear in r when w is small (it is 0.1 here). Of course, the population is not uniformly distributed, and as a result we see that the curves increase linearly only for a small distance. Beyond that the population density falls off and we see that the number of people falls off as well.

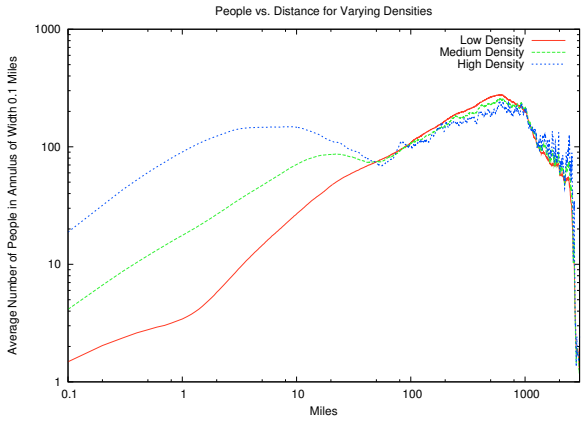

Figure 6: Number of individuals as a function of distance. Here we show how many people there are on average who live x miles away. We divide the US into low, medium and high density areas, and compute the curves independently for each.

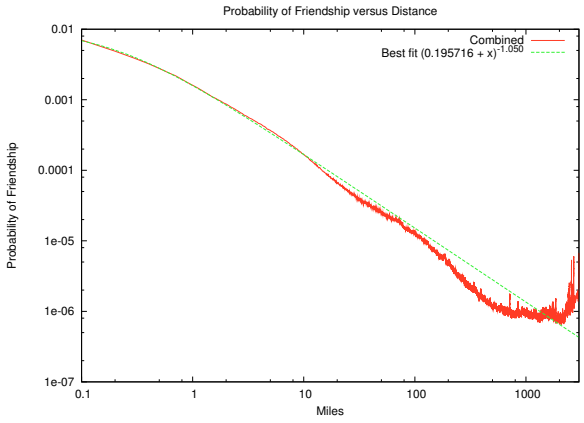

Figure 7: Probability of friendship as a function of distance. By computing the number of pairs of individuals at varying distances, along with the number of friends at those distances, we are able to compute the probability of two people at distance d knowing each other. We see here that it is a reasonably good fit to a power-law with exponent near -1.

This is caused by two competing forces: as we increase the radius, the area of the annulus increases, increasing the population we would expect to find. On the other hand, as we move further away from urban centers, we are more likely to find ourselves in the country, where population is sparse. At some point (about 50 miles) the annulus becomes sufficiently large such that it incorporates a wide swath where the average population density is quite unrelated to the density at the center of the annulus, and becomes more closely related to the average population density in the US. This causes the three curves to meet and overlap from 50 miles onward.

3.2 Friendship and Distance

We now turn to an investigation of the probability of friendship as a function of distance. Naturally, we expect the probability to go down with distance and this is what we observe in Figure 7. To generate this curve, we aggregate over all individuals, computing the distance between all $8.1 * 10^{12}$ pairs of individuals with known addresses. We then bucket by intervals of 0.1 miles to compute the total number of pairs and the number of pairs for which an edge is present, plotting the ratio. It turns out that we can get a good fit to a curve of the form $a(b+x)^{-c}$. The exponent very close to $c = -1$ suggests that, at medium to long-range distances, the probability of friendship is roughly inversely proportional to distance. At shorter scales the curve is flatter, suggesting that there is less sensitivity to short distances than a power-law with exponent -1 would produce. The -1 exponent has been observed in other datasets as well [15], suggesting that there is a more general principle at work here.

However, this does not tell the full story, as it aggregates people together from very different settings. When we break it down by population density in Figure 8, a somewhat different account emerges; for short distances the probability is higher in lower density areas as you are more likely to be friends with a person a few miles away if you live in a less dense area. Interestingly, as the distance increases, the three curves converge. At about 50 miles, we see that the probability of knowing someone is no longer dependent on density. In fact, as we go further away, the order inverts, with people in high density areas being more likely to be friends with people at greater distances. This supports the intuition that people living in metropolitan areas are more cosmopolitan; their ties to distant places are more likely, probably arising from increased movement between cities and greater capacity to travel.

An alternative to observing friendship probability as a function of distance is to look instead as a function of rank. As described in Liben-Nowell et al., we define rank as the number of people who live closer than a user. For user u, we rank users by distance from u. For user v, the number of people living in the area between u and v is defined by $rank_u(v) := |\{w : d(u,w) < d(u,v)\}|$. The hope here is that despite the differences in population density, the probability of being friends with someone at a given rank should be independent of where you live.

Figure 9 shows friendship probability as a function of rank. Here we do see a nice smooth curve, again with an exponent of close to -1 (as previously observed). Even though using rank should mitigate the effect of density on our probability calculation, it does not control for the behaviors of users in different areas. Figure 10 shows the probability of friendship as a function of rank, this time broken down by our three density groups. Though the curves do overlap somewhat more when we calculate things this way (all with exponent about -1), we still see similar effects. The probability is higher at low ranks for people in less dense areas, and higher at high ranks for people in more dense areas (cosmopolitan effect). This reinforces the notion that people who live in urban areas tend to have more dispersed friends.

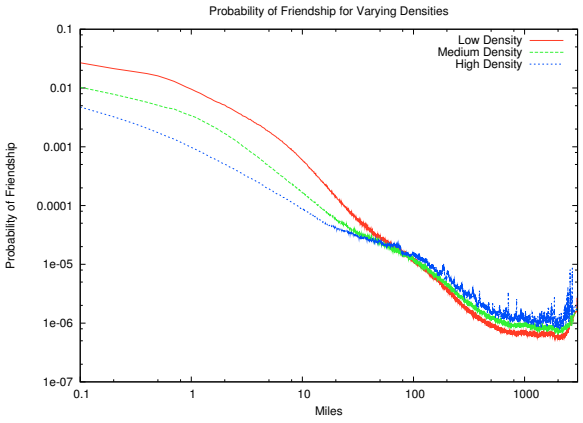

Figure 8: Looking at the people living in low, medium and high density regions separately, we see that if you live in a high density region (a city), you are less likely to know a nearby individual, since there are so many of them. However, you are more likely to have contact with someone far away.

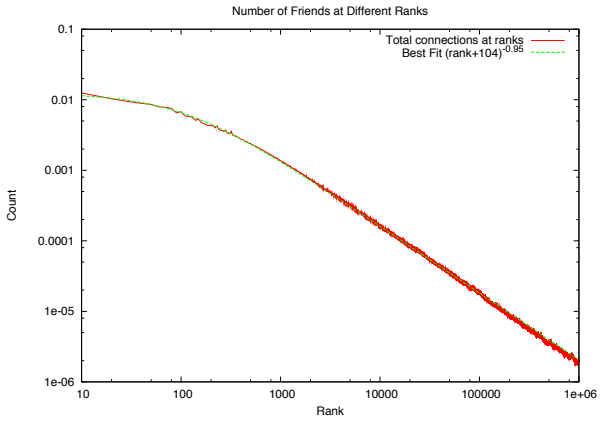

Figure 9: The rank of a person v relative to u is the number of individuals w such that $d(u,w) < d(u,v)$. Here we show the probability of friendship as a function of rank.

4. PREDICTING LOCATION

A practical application of the observations made thus far is that they allow us to predict the locations of people who have not provided this information. If we can accurately predict an individual's location, we can improve services for them in a number of ways. The most obvious application is that we can provide them with better local content. Providing a more local, personalized experience is likely to make a site more useful for users. We can also use a person's location to help prevent security breaches – if an individual accesses the site from a location far from home (where the individual's current location is approximated via IP geolocation), and they have never been there before, we might ask them an additional security question to ensure that their account has not been compromised. Thus, our goal here is, given the locations of a user's contacts, to compute that user's home location.

In the simplest case, all of one's friends would live in a small region, and then the prediction task would be very simple, with any reasonable algorithm returning a good approximation. Things get more complicated and difficult as one's friends become more spread out. The distributions from the previous sections tell us that one will typically not have too many friends at great distances, but that there will be too many for naive algorithms to work well.

For instance, a first attempt would be to take the mean location of one's friends. However, this will give laughably bad results for people living on either coast. An individual with 10 friends in San Francisco and one friend in New York will be placed an eleventh of the way from San Francisco to New York, somewhere in Nevada. Other simple statistics, like median (whatever that would mean in two dimensions) do better, but still fail, especially for people living on the coasts.

To achieve better performance, we must develop a more sophisticated model using the observations from the proceeding sections. Figure 7 shows the probability of an edge being present as a function of distance, which suggests a maximum likelihood approach. We consider an individual u with k friends. Using the distribution from Figure 7, we can computed the likelihood of a given location $l_u = (lat, long)$. For each friend v of u whose location l_v is known, we can compute the probability of the edge being present given the distance between u and v, $p(|l_u - l_v|) = 0.0019(|l_u - l_v| + 0.196)^{-1.05}$, as empirically determined.

Multiplying these probabilities together for all such v, we obtain a likelihood for all the edges. To complete the calculation, we must also multiply the probabilities of all the other edges not being present: $1 - p(|l_u - l_v|)$ for all v such that $v \notin E$. Because all of the probabilities are very small for any particular edge, this term serves mostly as a tiebreaker and typically plays a small role. Thus, we can write down the likelihood of a particular location l_u as

$$\prod_{(u,v)\in E} p(|l_u - l_v|) \prod_{(u,v)\notin E} 1 - p(|l_u - l_v|)$$

This model gives us a way to evaluate any point l_u. From a practical point of view, however, the algorithm as stated is very expensive. In a naive implementation, to find the best location for one individual, we would have to compute the probability terms for every other user, at an expense of $O(N)$ per location evaluated. Finding the best location would require an additional search, making this impractical in a large graph.

With two optimizations, however, we can develop an efficient algorithm which computes the (near) optimal locations for all individuals in $O(M \log N)$ assuming that the maximum degree in the graph is $O(\log N)$ (where M is the number of edges and N is the number of users).

The first important observation is that, for any location, the second part of the product, containing $1 - p(\cdot)$, is very

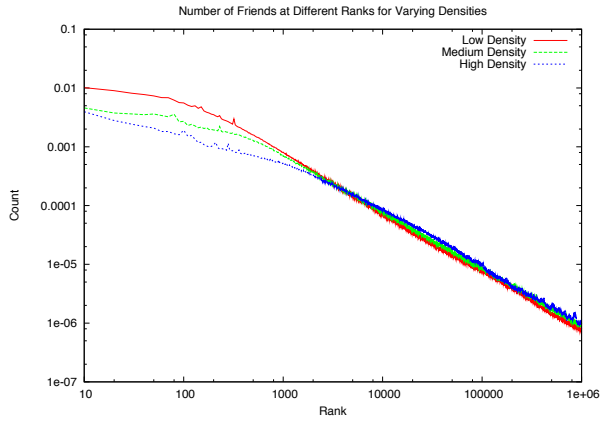

Figure 10: Similar to probabilty versus distance, here we see that people in higher density regions are less likely to know the low rank people living near them, but more likely to know the higher rank people living further away.

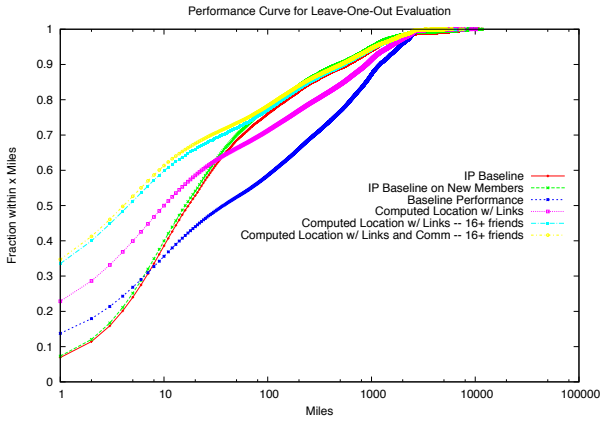

Figure 11: Location Prediction Performance. This figure compares external predictions from an IP geolocation service, the same service constrained to users who have recently updated their address, a baseline of randomly choosing the location of a friend, along with three predictions: our algorithm with all links, for users with 16+ friends, and finally for users with 16+ friends constraining to only those with whom they have communicated recently.

nearly independent of u. Thus, we can precompute a constant $\gamma_l = \prod_{v \in V} 1 - p(|l_u - l_v|)$ for each location l. We can then rewrite the above formula as:

$$\gamma_{l_u} = \prod_{(u,v) \in E} \frac{p(|l_u - l_v|)}{1 - p(|l_u - l_v|)}$$

The other important optimization comes from the form of the function $p(\cdot)$. This function is very sharply peaked at $p(0)$, and as a result the most likely location is typically colocated with one of u's friends.

In fact, if we ignore the γ term, we can prove that u would be colocated with a friend v if people lived in one dimension instead of two.

For a contradiction, imagine that $l_u \neq l_v$ for any friend of u. Then, the probability function in one dimension for a location x is $P(x) = \prod_{(u,v) \in E} (|x - x_v| + b)^{-c}$, for some positive constants b and c, where v is located at x_v. This function will have minima and maxima at the same locations if we log-transform it to get the more manageable equation $\sum_{(u,v) \in E} -c \log(|x - x_v| + b)$. We can split this up in to two terms, those where $x > x_v$ and those where $x < x_v$, yielding

$$\sum_{(u,v) \in E | x_v < x} \log(x - x_v + b) + \sum_{(u,v) \in E | x_v > x} \log(x_v - x + b)$$

When we take the second derivative and collect terms, we end up with $\sum_{(u,v) \in E} c(x - x_v + b)^{-2}$, which is always positive. Thus, there are can be no interior maxima, and the likelihood function is thus maximized at some x_v, where the derivative is undefined.

While this is not the case in two dimensions, and cases can be constructed where the maxima is not colocated with a friend, the one-dimensional analysis suggests that in many cases the maxima will be colocated with a friend. When we perform an exhaustive search of the two dimensional space, we find that in practice, the likelihood is almost always maximized at the location of a friend. It takes a contrived example to force the maxima somewhere other than a location very near some friend.

This allows us to greatly prune the geographic search space. Thus, to compute the most likely locations for a large group of users, our algorithm performs two steps. First, it precomputes γ for all locations (where all locations is a fine mesh of locations in the US). This is an expensive operation, but can be easily parallelized and must only be run once. Next, to make a prediction for an individual u, we evaluate the likelihood of all the locations of the friends of u, picking the best one. Thus, if u has k friends, the algorithm takes $O(k^2)$ to compute $p(\cdot)$ for all k friends from k locations. Since k is typically small, on the order of dozens, this is fast, and can also be easily parallelized. As a final note, it is important to do all the calculations adding logarithms instead of multiplying probabilities to avoid underflow.

4.1 Performance Methodology

To compute the performance of our algorithms, we take the provided address of the 2.9 millions users for whom we can obtain precise location as the ground truth. Naturally, some of these addresses are incorrect or out of date, but we believe that the vast majority of them are accurate. To quantify this, we find that 57.2% of users have IP addresses that geolocate to within 25 miles of their provided address. We compare this to those users who have updated their location within the last 90 days. If a significant fraction of the users had moved since last updating their addresses, we would expect IP geolocation to do significantly better on the users who had updated their address in the last 90 days, as the new addresses would be much more likely to be accurate. However, we find that the fraction correctly placed within 25 miles only increases to 58.5%.

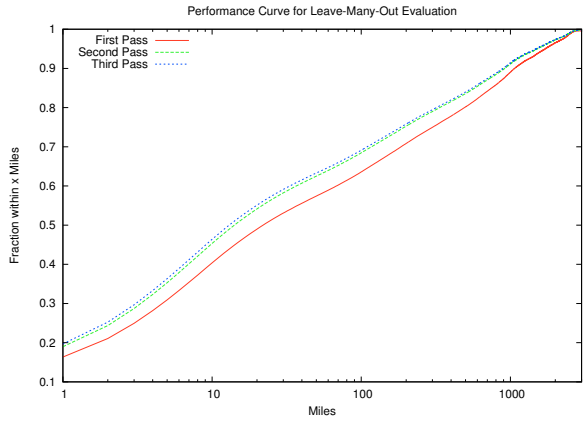

Figure 12: When we are predicting the locations of many individuals at once, we can perform better by using the information contained in the links between the individuals whose locations we are trying to predict. On the first pass, we make our prediction based only on the known addresses. On subsequent passes, we use the predicted locations as part of the input, improving performance.

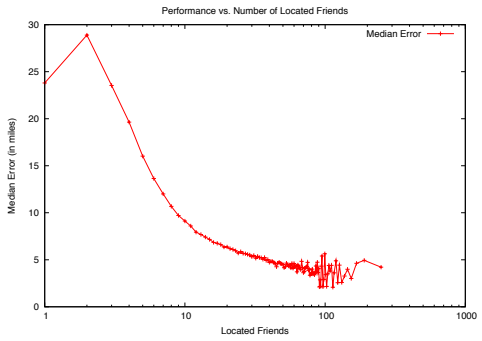

Figure 13: Prediction performance as a function of friend count. As friend count increases, more information allows for better geolocation

4.2 Leave-One-Out Evaluation

Figure 11 shows the performance of the maximum likelihood algorithm. To evaluate the algorithm, we predict the location of all 2.9 million users whose location is known, and who have at least one friend whose location is also known. For each user, we make our prediction based on the user's friends and then compare it to the location they provide. The figure shows, for instance, that we guess within 25 miles for 67.5% of the users with 16 or more located friends (the value 16 was chosen arbitrarily to illustrate that we do best with a moderate number of located friends). This compares favorably to other methods; in particular it does better than IP-based geolocation (57.2%), and performs much better than a baseline algorithm that picks a friend at random and colocates users to that location (46.3%). When comparing to the entire 2.9 million users, IP geolocation places a higher percentage of people within intermediate distances. For instance, IP geolocation is within 50 miles 68.4% of the time, while our algorithm only places 67.6% correctly. Most of this advantage comes from low-degree individuals, and when we look only at those with 16 or more friends, we do better than IP-based methods at all distances.

Overall, friend-based geolocation seems to be better than IP-based geolocation, so long as an individual has a sufficient number of friends. To improve performance further, we can use additional sources of information. The yellow line in Figure 11 creates a single extra edge between individuals who have communicated or viewed each other's profiles in the last 90 days. This places extra weight on some edges while creating a few others that are not explicitly present in the friendship graph. This gives us a performance boost from 67.5% to 69.1% (at 25 miles) on individuals with 16 or more (explicit) friendships.

Another approach is to use the probability versus rank function to make our predictions instead of the probability versus distance function. This approach is more computationally expensive because computing rank requires knowing how many people are closer and further than a given friend. However, rank can be approximating by sampling. Unfortunately, this approach seems to give no increase in performance when compared to the methods described above.

4.3 Leave-Many-Out Evaluation

Another evaluation method, and one which is more similar to the envisioned use cases, is to attempt to recover the locations of many individuals simultaneously. To do this, we remove the addresses from 75% of the individuals who have provided this information. We then attempt to recover the locations of all users who still have at least one friend remaining in the set with known addresses. In doing things this way, we are attempting to predict the addresses of 1.6 million users based on the addresses of 700,000 other users.

A first attempt at this is to simply run the algorithm from the proceeding section 700,000 times. However, this omits all of the information in the edges between the 1.6 million users for whom we are trying to locate. The performance curve shown in Figure 12 is much worse, as users now have only about one quarter as many geolocated friends for the prediction to be based on. Predicting in this way correctly places only 51.3% of users within 25 miles of their provided locations.

Ideally, we would place all of the individuals in such a way that we optimize the global likelihood, including the edges between two users of unknown location, and the edges between an unknown location and a known location. Unfortunately, we do not know how to do this in an efficient way.

However, that does not mean that we should throw away the information in the unknown to unknown edges. Instead, we can run our prediction algorithm iteratively, using the newly guessed locations as input as well as the locations provided by users.

Figure 12 shows the performance of this iterative approach. The second iteration is significantly better that the first (56.5% vs. 51.3% at 25 miles), and the third is slightly bet-

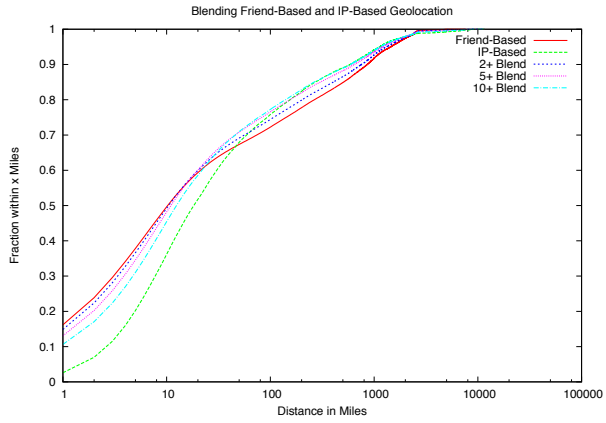

Figure 14: The accuracy of friend-based geolocation depends somewhat on how many located friends an individual has. By using IP-based geolocation for those with few friends, and friend-based geolocation for those with many friends, we can do better than either approach individually. Here we show the curves for just friend-based, just IP-based, and using friend-based for those with 2+, 5+, or 10+ friends while using IP for the rest.

ter than the second (57.4% vs. 56.5% at 25 miles). Beyond that, there is little improvement.

4.4 Combining friend and IP predictions

As a final evaluation, we would like to integrate all of our information sources to produce the best prediction possible for a given user. Figure 13 shows the median prediction error as a function of the number of geolocated friends. As one would expect, with more information from more friends, we are better able to predict the correct location of an individual.

This also suggests a way to improve our prediction performance. We expect that the quality of IP-based geolocation is independent of the number of friends a person has. Also, we know that friend-based geolocation works best on people with more friends. Thus, a simple way to combine these two predictions is to use IP-based geolocation on individuals with just a few friends, and use the friend-based geolocation on individuals with more friends. Figure 14 shows the performance as we vary the threshold. As the threshold increases, and more people are predicted based on IP, the fraction located within just a few miles drops, but the fraction located correctly within 100 miles increases.

Based on these results, it seems that a good tradeoff is to predict from friend locations when an individual has 5 or more locatable friends, and from the user's IP address if she has fewer than 5 friends with known addresses. Doing this causes the performance at 100 miles to slightly exceed the IP performance, and it is almost as good as strictly friend-based prediction at smaller distances.

5. CONCLUSIONS

Our examination of user-contributed address and association data from Facebook shows that the addition of social information to the task of predicting physical location produces measurable improvement in accuracy when compared to standard IP-based methods.

In this paper, we first analyze friendship as a function of distance and rank and generate several observations regarding the interplay of geography and friendship. We find that at medium to long-range distances, the probability of friendship is roughly proportional to the inverse of distance. However, at shorter ranges, distance does not play as large of a role in the likelihood of friendship. We also look at friendship probability as a function of rank (where rank is the number of people who live closer than a friend ranked by distance), and note that in general, people who live in cities tend to have friends that are more scattered throughout the country.

We then present an algorithm to predict the physical location of a user, given the known location of her friends. We find that using a maximum likelihood approach with the simplifying assumption that the user will be either colocated or in close proximity to one of her friends, we are able to guess the physical location of 69.1% of the users with 16 or more located friends to within 25 miles, compared to only 57.2% using IP-based methods. We then investigate how even more social data may further improve geolocation results, using data on how often users interact with each other and see each other's content. Using this data generates slight improvement in geolocation, which implies that users who are physically close to each other may tend to interact more often on Facebook.

We also embark on a more ambitious effort to predict the location of many individuals at once. Iterating our maximum-likelihood algorithm provides significant improvement in the accuracy of our predictions.

Having more accurate data of a user's physical location would improve efforts to predict new friendships and associations (which in turn improves the friend suggestion tool). However, there are many other applications as well. For example, algorithms to detect adversarial account takeovers would be improved with better location data of a particular user and her friends. Socially predicted locations could also be used to calibrate and verify other geolocation data, such as latitude/longitude information contained in EXIF metadata from photos. We could even use these methods in an attempt to improve the IP-to-location conversion process.

Iteration of our algorithms would allow us to derive location predictions for the majority of users who have not yet provided address information. This has clear applications for the provision of location-based services.

Future work may further improve precision in our quest to obtain the best possible location prediction for a particular user. In addition to using edge data from the social graph, we may supplement our data using social events as a proxy to coincident location. For example, we can infer closeness between two individuals if we observe a photo tagged with both users, colocating them at a point in time. Events attended by two or more individuals may also provide useful data, especially if an address is provided for the event. It may also be beneficial to attach timestamps to all of our data sources and weight these signals appropriately when predicting a user's current location. We expect, for instance, that

newly formed relationships should have more weight than old ones, as new relationships are more likely to be formed at one's current address, whereas an older relationship could be, for instance, an old friend from high school.

Finally, while location lookup based on IP address is quite well-developed in the US, the accuracy is much worse in some countries. Though we only evaluated our methods on US users, we expect that these results will be internationally applicable and will allow us to improve our location estimates in countries where IP address often tells no more than the name of the country.

6. ACKNOWLEDGEMENTS

We would like to thank Stephen Heise for his work on building a geocoder service that made our experimentation possible.

7. REFERENCES

[1] L. Adamic, R. Lukose, A. Puniyani, and B. Huberman. Search in power-law networks. *Physical review E*, 64(4):46135, 2001.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, New York, NY, USA, 2006. ACM Press.

[3] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 357–366, New York, NY, USA, 2008. ACM.

[4] C. Butts. Predictability of large-scale spatially embedded networks. In *Dynamic Social Network Modeling and Analysis: workshop summary and papers*, pages 313–323, 2003.

[5] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *WWW*, 2009.

[6] R. Dhamija, J. D. Tygar, and M. Hearst. Why phishing works. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 581–590, New York, NY, USA, 2006. ACM.

[7] L. Festinger, S. Schachter, and K. Back. *Social pressures in informal groups: A study of human factors in housing*. Stanford Univ Pr, 1963.

[8] E. Gilbert, K. Karahalios, and C. Sandvig. The network in the garden: an empirical analysis of social media in rural life. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1603–1612, New York, NY, USA, 2008. ACM.

[9] S. Graham. The end of geography or the explosion of place? Conceptualizing space, place and information technology. *Progress in human geography*, 22(2):165, 1998.

[10] A. Khalil and K. Connelly. Context-aware telephony: privacy preferences and sharing patterns. In *CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 469–478, New York, NY, USA, 2006. ACM.

[11] J. Kleinberg. Navigation in a small world. *Nature*, 406(6798):845–845, 2000.

[12] J. Kleinberg. The small-world phenomenon: an algorithm perspective. In *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170, New York, NY, USA, 2000. ACM Press.

[13] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 617. ACM, 2006.

[14] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470. ACM, 2008.

[15] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences*, 102(33):11623, 2005.

[16] MaxMind, Inc. GeoIP City Accuracy for Selected Countries, November 2008. `http://www.maxmind.com/app/city_accuracy`.

[17] B. Mayhew and R. Levinger. Size and the density of interaction in human aggregates. *The American Journal of Sociology*, 82(1):86–110, 1976.

[18] D. Mok and B. Wellman. Did distance matter before the Internet? Interpersonal contact and support in the 1970s. *Social networks*, 29(3):430–461, 2007.

[19] L. Nahemow and M. Lawton. Similarity and propinquity in friendship formation. *Journal of Personality and Social Psychology*, 32(2):205–213, 1975.

[20] J. Q. Stewart. An inverse distance variation for certain social influences. *Science*, 93(2404):89–90, 1941.

[21] U.S. Census Bureau. Census 2000 Summary File 1, 2000. `http://factfinder.census.gov/servlet/DCGeoSelectServlet?ds_name=DEC_2000_SF1_U`.

[22] U.S. Census Bureau. Redistricting Census 2000 TIGER/Line Files, 2000. `http://www.census.gov/geo/www/tiger/tiger2k/tgr2000.html`.